УДК 519.7:007.52;519.81

О ПОДХОДАХ К ОЦЕНИВАНИЮ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ В ТЕСТОВОМ РАСПОЗНАВАНИИ

С.И. Колесникова

Томский государственный университет систем управления и радиоэлектроники
E-mail: skolesnikova@yandex.ru

Предлагаются два подхода к вычислению весовых коэффициентов характеристических признаков, используемых в тестовых системах поддержки принятия решения, а именно: подход на основе формализма мультимножеств и метода анализа иерархий и подход на основе упрощенного метода анализа иерархий, частично решающего проблему определения весовых коэффициентов признаков для случая, когда размерность признакового пространства достаточно велика. Обсуждаются методы, реализующие данные подходы.

Одной из наиболее важных проблем при создании интеллектуальных систем выявления закономерностей и поддержки принятия решений является проблема анализа признакового пространства на предмет выделения наиболее значимых признаков и оценивания величины их значимости [1, 2]. Несмотря на то, что этой проблеме посвящено большое количество публикаций, например, [2–4], до настоящего времени отсутствует корректное сравнение методов ее решения.

При решении многокритериальных задач выбора понятие относительной важности (весомости, значимости) критериев является основным [1–4]. Следует отметить, что более детально представлены результаты по многокритериальным функциям предпочтения, использующих базовые шкалы оценок альтернатив по критериям [5]. Функция предпочтения, как правило, представляет собой отображение множества альтернатив на числовую ось, и лучшей альтернативе приписывается большее число.

В интеллектуальных системах, основанных на методах тестового распознавания образов [2, 5, 7], для принятия решения используются «хорошие» тесты, т. е. тесты, содержащие меньшее количество признаков и с большим весом, где под весом теста понимается сумма весовых коэффициентов признаков [2].

Работа посвящена вычислению весовых коэффициентов характеристических признаков (ВКП), используемых в системах поддержки принятия решения [1, 2], основанных на тестовых методах распознавания образов.

Основные определения и понятия

Нижеприведенные методы вычисления весовых коэффициентов характеристических признаков основаны на матричной модели представления данных и знаний, включающей матрицу описаний Q объектов в пространстве характеристических

признаков и матрицу различений R объектов в пространстве классификационных признаков [2].

Элемент q_{ij} матрицы Q задает значение j-го признака для i-го объекта. Строка q_i матрицы Q сопоставляется объекту $s_i(i\in\{1, ..., l\})$, где l – число обучающих объектов. В случае, если у какого-либо объекта в матрице Q значение признака отмечено символом «–», то считается, что значение соответствующего характеристического признака безразлично (признак может принимать как нулевые, так и единичные значения, а в случае k-значных признаков – любые целочисленные значения признаков из заданного интервала значений признака).

Строки матрицы различений R сопоставляются одноименным строкам матрицы Q, столбцы – классификационным признакам, определяющим различные механизмы разбиения объектов на классы эквивалентности (механизмы классификации). Элемент r_{ij} матрицы R задает принадлежность i-го объекта одному из выделенных классов по j-му механизму классификации. Для указания факта принадлежности объекта классу используется номер этого класса при соответствующем механизме классификации. Множество всех неповторяющихся строк матрицы R сопоставлено множеству выделенных образов. Элементами образа являются объекты, представленные строками матрицы Q, сопоставленными одинаковым строкам матрицы R. Если имеется единственный механизм классификации, матрица различений вырождается в столбец, что соответствует традиционному представлению знаний в задачах распознавания образов [2].

Задача распознавания состоит в определении по матрицам Q и R образа, которому принадлежит заданный совокупностью признаков исследуемый объект, как правило, не входящий в обучающую выборку.

Назовем признаки зависимыми, если имеется хотя бы одна пара объектов из разных образов, различаемая этими признаками.

Совокупность признаков, различающих все пары объектов из разных образов (классов) при каждом механизме классификации, назовем диагностическим тестом (далее слово «диагностический» будем опускать).

Тест назовем безызбыточным (тупиковым [1]), если исключение любого признака из теста нарушает его свойство быть тестом.

При $q_{ij}\in\{0,1,-\}$ вводятся следующие определения [2].

1. Два объекта считаются различимыми, если хотя бы один характеристический признак в описании одного из них принимают значение 1 (0), а в описании другого – инверсное, т. е. значение 0 (1).
2. Под весовым коэффициентом признака понимается числовая оценка его различающей способности.

Постановка задачи. Пусть по матрицам Q и R построены все (часть) безызбыточные тесты, представленные матрицей тестов T, строки которой сопоставлены тестам, а столбцы – характеристическим признакам, и определено число различающих пар «объект-объект» по каждому характеристическому признаку.

Требуется определить весовые коэффициенты характеристических признаков, входящих в объединение всех (части) безызбыточных тестов [2]. При этом не исключается возможность достаточно большой размерности признакового пространства и наличия взаимозависимости признаков.

Методы оценивания весовых коэффициентов признаков

При вычислении весовых коэффициентов характеристических признаков в тестовом распознавании [2], как правило, требуется выполнение условия независимости признаков (критериев) по предпочтению, либо не учитывается реально возможная их взаимозависимость (в смысле влияния оценки различающей способности одного из них на оценку различающей способности другого).

Остановимся кратко на методах оценивания весовых коэффициентов признаков в интеллектуальных системах поддержки принятия решений, используемых для вычисления весов тестов и имеющих отношение к поставленной задаче.

1. В методе, основанном на различающей способности признака [2], предлагается следующая формула для вычисления весового коэффициента w_m m-го признака:

$$w_m = \frac{\sum_{r=1}^{K-1}\sum_{t=r+1}^{K}\sum_{i=1}^{N_r}\sum_{j=1}^{N_t}\delta_{ij}^m}{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\sigma_i\sigma_j}, \qquad (1)$$

где K – число выделенных образов; N_f – число строк в описании f-го образа; $f\in\{r,t\}$, $\delta_{ij}^m=0$, если $q_{im}=q_{jm}=0$ или $q_{im}=q_{jm}=1$ (q_{im} и q_{jm} – элементы матрицы Q из разных образов); $\delta_{ij}^m=p_ip_j2^{d_i^-+d_j^-}$; ($d_i^-$ – число значений «–» в i-ой строке матрицы Q, p_i – коэффициент повторения i-ой строки), если $q_{im}=0$ и $q_{jm}=1$ или $q_{im}=1$ и $q_{jm}=0$; $\delta_{ij}^m=p_ip_j2^{d_i^-+d_j^--1}$, если q_{im}=«–» и (или) q_{jm}=«–»; σ_j – число объектов в j-ом образе ($j=1,...,K$), вычисляемое по формуле:

$$\sigma_j = \sum_{k=1}^{N_j} p_k 2^{d_k^-};\ m\in\{1,2,...,M\}.$$

Метод вычисления весовых коэффициентов признаков на основе оценки его различающей способности признаков весьма прозрачен, прост и эффективен при репрезентативной выборке, но при условии независимости оцениваемых признаков.

2. Основанный на формализме мультимножеств [3] **и методе анализа иерархий** (МАИ) [4] метод определения ВКП учитывает вклад признаков в распознающую способность теста с учетом их взаимозависимости и базируется на представлении совокупности всех различимых пар объектов из разных образов для каждого признака z_i, $i=1,2,...,M$, в

виде мультимножества, применении метода анализа иерархий Саати, использующего парные сравнения признаков на основе специальным образом выбранной меры относительной важности признаков, учитывающей их взаимозависимость [5].

Соответствующий признаку z_m m-й столбец матрицы Q порождает совокупность P_m различимых этим признаком пар объектов из разных образов $\{(i–j)|i\in F_r, j\in F_t, r\neq t, q_{im}\neq q_{jm}\}$, где $F_l=\{f_l, f_l+1,...,f_l+N_l-1\}$, N_l – число строк в l-ом образе, f_l – номер 1-й строки в матрице Q для l-го образа, $l\in\{r,t\}$, т. е. признак z_m порождает мультимножество P_m.

Метод состоит из трех этапов, на каждом из которых формируется матрица парных сравнений (МПС) признаков $z_{i_1}, z_{i_2},...,z_{i_m},...,z_{i_g}$ ($i_m\in\{1, ..., M\}$, g – количество признаков в тесте) на основе определенной меры относительной важности признака i над признаком j, в качестве которой поэтапно выбираются величины:

$$\text{(I)}\ \frac{|P_i|}{|P_j|},\quad \text{(II)}\ \frac{\delta(|P_i-P_j|)}{\delta(|P_j-P_i|)},\quad \text{(III)}\ \frac{\delta(/P_i-P_j/)}{\delta(/P_j-P_i/)},\qquad (2)$$

где $|P_i|$ – мощность (общее количество элементов) i-го мультимножества, соответствующего признаку z_i, $/P_i/$ – размерность (количество элементов, встречающихся один раз) i-го мультимножества, $P_i–P_j$ – разность мультимножеств, соответствующих признакам z_i и z_j, где $\delta(x)=x$, если $x\neq0$, и $\delta(x)=1$, иначе.

На s-м ($s\in\{1,2,3\}$) этапе по данному методу вычисляется g – компонентный вектор значений весовых коэффициентов признаков – $W_s=w_{s,1}, w_{s,2},...,w_{s,q}$, совпадающий со значением нормализованного главного собственного вектора соответствующей матрицы (по Саати [4] – локальные приоритеты). Вектор $W=(W_1\cdot W_2\cdot W_3)^{1/3}$ представляет собой обобщенные значения весовых коэффициентов признаков (глобальные приоритеты [4]), входящих в тест.

Вышеуказанные меры позволяют учесть не только общие свойства сравниваемых признаков (степень сходства или различия), но, что особенно важно, и уникальные их свойства (степень приоритетности одного признака над другими).

Следующий результат устанавливает связь между двумя вышеупомянутыми методами.

Теорема 1. Нормализованный главный собственный вектор матрицы парных сравнений признаков $W_1=(w_{1,i_1}, w_{1,i_2},...,w_{1,i_n})$ с мерой относительной важности признака i над признаком j, равной $\frac{|P_i|}{|P_j|}$, равен вектору $W_0=(w^0_{i_1}, w^0_{i_2},...,w^0_{i_g})$, где w^0_m – нормализованный коэффициент (m, $i_m\in\{1, ..., M\}$), полученный по формуле (1).

Известно, что при использовании метода анализа иерархий указанный способ определения весового вектора существенно опирается на свойство совместности матрицы парных сравнений и не является обоснованным в случае его нарушения.

Решение задачи

Кратко изложим подход, основанный на формализме мультимножеств и упрощенном МАИ, который частично решает проблему нахождения ВКП для случая достаточно большой размерности признакового пространства. Как известно, данная задача до сих пор является нетривиальной не только в случае задания предпочтений экспертами вручную, но и в случае, когда относительная важность ВКП вычисляется по предложенным в [2, 5] формулам. Например, в случае 50-ти признаков, используемых в задачах диагностики, количество парных сравнений составит 1200, что является не только трудоемким процессом, но и проблематичным для хранения такого объема данных в памяти ЭВМ.

Данный подход существенно опирается на идею В.Д. Ногина [6] о «базисных» элементах (расположенных выше (либо ниже) главной диагонали), на основе которых затем легко и без ошибок вычислительного характера находится искомый весовой вектор. Выбор конкретного «базисного» набора соответствует той или иной схеме сравнения объектов, которую можно выбирать с целью получения наиболее надежных результатов. Методы, реализующие данный подход, во-первых, основаны на совместной матрице A и, таким образом, избавлены от «модельной» ошибки; во-вторых, обеспечивают количество требуемых сравнений существенно меньшее традиционно требуемых.

Для изложения метода перечислим требования к матрице относительных весов

$$\mathrm{A}=\left\|a_{ij}\right\|_{M\times M},\ a_{ij}=\frac{w_i}{w_j},$$

где w_i, w_j – компоненты весового вектора $W=(w_1,w_2,...,w_M)^T$ [4]: 1) $a_{ij}>0$; $i,j=1,...,M$; 2) $a_{ij}=a_{ij}^{-1}$; $i,j=1,...,M$; 3) $a_{ij}=a_{ik}\cdot a_{kj}$ $i,j,k=1,...,M$; 4) число M является максимальным собственным значением матрицы A, и для некоторого единственного (нормированного) вектор-столбца $W=(w_1,w_2,...,w_M)^T$ с положительными компонентами выполняется равенство: $A\ W=M\ W$.

Подход к решению данной проблемы, связанный с упрощенным вариантом МАИ, автором адаптирован следующим образом.

Сначала выбираются $M–1$ базисных элементов, на основе которых определяются все остальные элементы МПС. Способ выбора базисных элементов является задачей творческой и зависит от модели задачи (и решения эксперта), в частности, от выбранной меры сравнения относительной важности признаков. Применительно к рассматриваемой задаче рассмотрим два следующих метода в рамках указанного подхода.

Первый метод связан с выбором некоторого «идеального» признака-образца, в качестве которого может служить либо признак с наибольшим значением ВКП, рассчитанным по формуле (1), либо объединение $\bigcup_i P_i$ всех мультимножеств, по-

рожденных соответствующими признаками (i=1,...,M). Далее формируем первую строку на основе соответствующей меры относительной важности (2) признака-образца над остальными признаками, элементы которой являются базисными. Остальные элементы МПС находятся из условия, обеспечивающего выполнение свойства совместности матрицы парных сравнений:

$$a_{ij} = a_{i1} \cdot a_{1j} = \frac{a_{1j}}{a_{1i}}, \; i = \overline{2,M}, \; j = \overline{1,M}. \qquad (3)$$

Тогда компоненты весового ненормализованного вектора $W^У=(w_1^У,w_2^У,...,w_M^У)^T$ (верхний индекс «$У$» означает «упрощенный»), следуя теореме 1 [6], вычисляются по формуле:

$$w_i^У = \frac{a_{iM}}{a_{1i}}, \; i = \overline{1,M}. \qquad (4)$$

Таким образом, компоненты весового вектора составляют последний столбец МПС A, элементы которой получены по формуле (3).

Преимущество данного способа с точки зрения трудоемкости вычислений очевидно: вместо вычисления $M(M–1)/2$ величин потребуется вычислить только $2M–3$ величин, т. е. в $M/4$ раза меньше при достаточно большом объеме M. К тому же, в случае экспертного задания оценок, формулы (3), (4) обеспечивают не только экономию времени, но и согласованность МПС, т. е. избавляют от «модельной» ошибки.

Второй метод предполагает выбор базисных элементов $a_{12},a_{23},...,a_{M-1,M}$, который реализует схему последовательного сравнения (некоторый «идеальный образец» сравнивается с 1-м, затем 1-й сравнивается со вторым и т. д.).

Сформулируем основные результаты.

Теорема 2. Пусть МПС $A=\|a_{ij}\|_{M\times M}$ построена по следующим правилам:

1) элементы $a_{12},a_{23},...,a_{M-1,M}$ являются базисными и определяются по формуле:

$$a_{ij} = \frac{\delta(|P_i - P_j|)}{\delta(|P_j - P_i|)} \left(a_{ij} = \frac{\delta(/P_i - P_j/)}{\delta(/P_j - P_i/)} \right), \qquad (5)$$
$$i = \overline{1,M-1}, \; j = \overline{2,M};$$

2) остальные элементы определяются по формуле:

$$a_{ij} = a_{i,j-1} \cdot a_{j-1,j}, \; i = \overline{1,M-2}, \; i < j-1. \qquad (6)$$

Тогда МПС $A=\|a_{ij}\|_{M\times M}$ удовлетворяет всем требованиям относительных весов 1–4, и компоненты ненормированного собственного вектора $W^У$ определяются по формуле:

$$w_i = a_{i,i+1} \cdot a_{i+1,i+2} \cdots a_{M-1,M}, \; i = \overline{1,M-1}, \; w_M = 1.$$

Теорема 3. Если матрица парных сравнений $A=\|a_{ij}\|_{M\times M}$ с элементами, вычисленными по формуле (2), удовлетворяет всем требованиям относительных весов 1–4, то ее собственный вектор совпадает с собственным вектором МПС, элементы которой вычислены по формулам (5), (6).

Отметим, что второй способ предполагает вычисление компонентов собственного вектора по базисным элементам, определяемым по точным формулам (2), и количество требуемых сравнений определится числом $M–1$, что в $M/2$ раз меньше традиционно требуемых сравнений.

Иллюстративный пример. Пусть заданы матрицы Q, R, T, представленные на рисунке. Вычислим для отдельного теста значения весовых коэффициентов признаков по вышеизложенным методам, например, для теста $T_1=(z_5,z_9,z_{11})$.

Для простоты изложения матрица различений R представлена одним столбцом.

Заметим, что число строк матрицы Q увеличится на 3, так как каждая из строк 3, 6, 7 представляет собой интервал булева пространства (содержит по одному символу «–») и представляется двумя строками.

Сравним численные значения весовых коэффициентов признаков, полученных по методу на основе МАИ [5] и методу, использующему упрощенный МАИ. Ввиду громоздкости МПС при большой ее размерности, а также имея в виду, что интерес, главным образом, представляет: а) погрешность в точности получаемых оценок по данным методам, т. е. $|W^M–W^У|$, где W^M и $W^У$ – нормализованные собственные векторы, соответствующим МПС; б) факт совместности МПС или значение индекса согласованности в указанных методах, рассмотрим получение значений весовых коэффи-

Q =

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	1	1	1	1	1	1	1	1	1	1	1	0
2	1	1	0	0	0	1	0	1	0	1	1	0	0
3	1	1	1	1	0	–	0	0	0	1	1	0	0
4	1	0	1	0	0	1	0	0	0	0	1	0	0
5	1	0	1	1	0	1	1	0	1	0	1	1	0
6	1	–	1	1	0	1	1	1	1	1	0	1	0
7	1	0	0	0	1	0	1	1	1	1	0	–	0
8	1	1	0	0	1	0	1	1	0	0	1	0	1

; R =

1
2
2
2
3
4
5
6

; T =

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	0	0	1	0	0	0	1	0	1	0	0
2	0	0	0	0	1	0	0	0	0	0	1	1	0
3	0	0	0	0	1	1	1	0	0	0	1	0	0
4	0	0	1	0	1	0	1	0	0	0	1	0	0
5	0	0	0	1	1	0	1	0	0	0	1	0	0
6	0	0	0	0	1	0	1	0	0	1	1	0	0

Рисунок. *Матрицы описаний Q, различений R и тестов T*

циентов признаков для вышеупомянутого теста $T_1=(z_5,z_9,z_{11})$, построенного по матрицам Q, R, T, рисунок.

В матрице парных сравнений признаков на основе меры относительной важности $\frac{\delta(|P_i-P_j|)}{\delta(|P_j-P_i|)}$ расположим признаки по неубыванию мощностей соответствующих разностей мультимножеств (табл. 1), учитывая, что $|P_5-P_9|=13$, $|P_9-P_5|=15$, $|P_5-P_{11}|=14$, $|P_{11}-P_5|=14$, $|P_9-P_{11}|=10$, $|P_{11}-P_9|=8$.

Таблица 1. *МПС признаков на основе меры относительной важности II и МАИ*

	P_9	P_5	P_{11}	W^M
P_9	1	15/13	5/4	0,375
P_5	13/15	1	14/14	0,317
P_{11}	4/5	14/14	1	0,308

МПС признаков на основе выбранной меры относительной важности и упрощенного МАИ, т. е. формул (5), (6), имеет вид (табл. 2):

Таблица 2. *МПС признаков на основе меры относительной важности II и упрощенного МАИ*

	P_9	P_5	P_{11}	W^Y	$\|W^M-W^Y\|$
P_9	1	15/13	5/4	0,375	0,000
P_5	13/15	1	(5·13)/(4·15)	0,325	0,008
P_{11}	4/5	(4·15)/(5·13)	1	0,300	0,008

Обратим также внимание на погрешность, связанную с индексом совместности МПС по методу, основанному на мультимножествах и МАИ, имея в виду, что уровень отношения согласованности равен 0,001 (для применимости МАИ желательным считается значение, меньшее 0,1 [4]):

$$\left\|\left|a_{ij}-\frac{w_i}{w_j}\right|\right\| = \begin{bmatrix} 0 & 0,031 & 0,033 \\ 0,023 & 0 & 0,027 \\ 0,022 & 0,026 & 0 \end{bmatrix}.$$

Таким образом, «модельная ошибка» в методе, основанном на мультимножествах и точных формулах (1), оказалась больше (максимальное отклонение равно 0,033), чем максимальное отклонение $|W^M-W^Y|$, равное 0,008, позволяющих в 2 раза сократить количество сравнений и упростить процедуру вычисления собственного вектора.

Для сравнительной оценки двух изложенных подходов к определению весов тестов может быть использован функционал качества, аналогичный введенному в [7], минимум которого укажет на более эффективный подход при принятии итогового решения для конкретной задачи распознавания.

Заключение

Весьма важным моментом в процессе выявления закономерностей данных и принятия решений в интеллектуальных системах является не только определение совокупности наиболее информативных (значимых) признаков (критериев), но и выявление отношений между ними. Приведенные методы позволяют количественно учесть вклад отдельного признака в различающую способность теста в условиях независимости (формула (1)) и возможной взаимозависимости (формулы (2), (5), (6)) входящих в тест признаков в условиях большой размерности признакового пространства.

Иллюстративный пример дает основание полагать, что учет специфики задачи с целью выбора соответствующего (более подходящего) метода приводит к построению более точных оценок значений весовых коэффициентов признаков, используемых в интеллектуальных системах поддержки принятия решений различного назначения.

Изложенный подход к определению ВКП, основанный на мультимножествах и упрощенном МАИ, может быть применен для решения реальных задач большой размерности с динамическими предпочтениями и приоритетами. Так, задача прогнозирования экспертных предпочтений связана с получением оценок приоритетности альтернатив в форме зависимостей от времени.

Для динамических задач матрица парных сравнений содержит функции времени в качестве элементов, поэтому максимальное собственное число λ_{max}, а также собственный вектор W также будут зависеть от времени, т. е. $A(t)\ W(t)=\lambda_{max}(t)\ W(t)$. Известно, что для этого уравнения можно получить аналитическое решение [4], если порядок матрицы $A(t)$ не превышает четырех. В противном случае данное уравнение решается численными методами для различных моментов времени t с последующей аппроксимацией с целью получения зависимости компонент вектора приоритетов $W(t)$ от времени. Предложенный метод вычисления ВКП на основе упрощенного МАИ позволяет снять ограничение на порядок матрицы $A(t)$ и следить за ее согласованностью во времени.

Теорема 1 содержит обоснование подхода, позволяющего существенно сократить количество вычислений при подсчете весовых коэффициентов признаков с мерами относительной важности признака i над признаком j, см. (2), учитывающими возможную взаимозависимость входящих в тест признаков.

Теоремы 2 и 3 содержат условие, гарантирующее применение предложенного подхода в задачах тестового распознавания без существенной погрешности.

Работа выполнена при финансовой поддержке РФФИ (проект № 04-01-00144), РГНФ (проект № 06-06-12603в).

СПИСОК ЛИТЕРАТУРЫ

1. Журавлев Ю.И., Гуревич И.Б. Распознавание образов и анализ изображений // Искусственный интеллект в 3-х кн. Кн. 2. Модели и методы: Справочник / Под ред. Д.А. Поспелова. – М: Радио и связь, 1990. – С. 149–190.
2. Yankovskaya A.E. Test Pattern Recognition with the Use of Genetic Algorithms // Pattern Recognition and Image Analysis. – 1999. – V. 9. – №. 1. – P. 121–123.
3. Петровский А.Б. Упорядочивание и классификация объектов с противоречивыми признаками // Новости искусственного интеллекта. – 2003. – № 4. – С. 34–43.
4. Саати Т.Л. Принятие решений. Метод анализа иерархий. – М.: Радио и связь, 1993. – 315 с.
5. Янковская А.Е., Колесникова С.И. О применении мультимножеств к задаче вычисления весовых коэффициентов признаков в интеллектуальных распознающих системах // Новости искусственного интеллекта. – 2004. – № 2. – С. 216–220.
6. Ногин В.Д. Упрощенный вариант метода анализа иерархий на основе нелинейной свертки критериев // Журнал вычислительной математики и математической физики. – 2004. – Т. 44. – № 7. – С. 1259–1268.
7. Янковская А.Е., Колесникова С.И. Поддержка принятия решений, коллективная оценка весовых признаков в интеллектуальных системах // Интеллектуальные системы. Интеллектуальные САПР: Труды Междунар. научно-техн. конференций. – М.: Физматлит, 2004. – С. 249–255.

УДК 519.6:004.652.4